DOCUMENT RESUME

ED 441 404 IR 057 675

AUTHOR	Aliprand, Joan M.
TITLE	Cataloguing in the Universal Character Set Environment: Looking at the Limits.
PUB DATE	1999-08-00
NOTE	9p.; In: IFLA Council and General Conference. Conference Programme and Proceedings (65th, Bangkok, Thailand, August 20-28, 1999); see IR 057 674.
AVAILABLE FROM	For full text: http://www.ifla.org/IV/ifla65/papers/079-155e.htm.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Alphabets; *Bibliographic Records; *Cataloging; Second Languages; *Standards
IDENTIFIERS	Anglo American Cataloging Rules 2 Revised; MARC; *Transcription; *Unicode

ABSTRACT

A new era for multilingual, multiscript computing is beginning, due to the development of the Unicode Standard and International Standard ISO/IEC 10646. The character content in these publications is kept carefully synchronized. With the addition of Ethiopic, Mongolian, and Sinhala, all of the world's major scripts are covered. Catalogers may expect that such an extensive character repertoire will meet all their needs for exact transcription of bibliographic data. This paper examines the topic of exact transcription and situations where it is not applied currently. The conceptual structure underpinning the character repertoire of the Unicode Standard and ISO/IEC 10646 is explained, followed by a discussion of whether the use of simple strings of characters can meet all needs for exact transcription. (Contains 15 references.) (Author/MES)

IFLANET Search Contacts
International Federation of Library Associations and Institutions
Annual Conference

65th IFLA Council and General Conference

Bangkok, Thailand, August 20 - August 28, 1999

Conference Proceedings

Code Number: 079-155(WS)-E
Division Number: IV
Professional Group: Cataloguing: Workshop
Joint Meeting with: -
Meeting Number:
Simultaneous Interpretation: *No*

Cataloguing in the universal character set environment: looking at the limits

Joan M. Aliprand
Senior Analyst, The Research Libraries Group
Mountain View, California, USA

Abstract

A new era for multilingual, multiscript computing is beginning, due to the development of the Unicode Standard and International Standard ISO/IEC 10646. The character content in these publications is kept carefully synchronized. A major milestone has now been reached. With the addition of Ethiopic, Mongolian and Sinhala, all of the world's major scripts are covered.

Cataloguers may expect that such an extensive character repertoire will meet all their needs for exact transcription of bibliographic data. This paper examines the topic of exact transcription, and situations where it is not applied currently. The conceptual structure underpinning the character repertoire of the Unicode Standard and ISO/IEC 10646 is explained, followed by a discussion of whether the use of simple strings of characters can meet all needs for exact transcription.

Paper

My first job was as a cataloguer, and though I'm now a systems analyst, I've maintained an active interest in the field. When I was learning about cataloguing in library school, the first edition of the *Anglo-American Cataloguing Rules*, the first rules based on the International Cataloguing Principles, was about to be published. I though that this was the last word on cataloguing, and not much more could be said. How wrong I was! And little did I dream that I would be contributing to the ongoing dialogue.

The focus of my presentation is descriptive cataloguing; chiefly the items that used to be called the "body" of the entry. Although I am focussing on descriptive cataloguing, some of what I say may be applicable generally, i.e., to all parts of bibliographic records, and even to other types of library records.

In my presentation, I'll refer to AACR2.[1] Now, I know that AACR2 isn't used everywhere. However, because I come from an English-speaking environment, these are the rules I know about. In addition, AACR2 has had an unusually broad influence: both directly and indirectly. Its direct influence has been through translations into other languages to serve as the basis for other cataloguing rules. It has indirect influence whenever one of the very large number of records created in the English-speaking world is used for copy cataloguing. Even when English is not the language of cataloguing, the information transcribed from the source of information might be useful and save time.

Rule 1.0E of AACR2, *Language and script of the description*, states in part:

> In the following areas, give information transcribed from the item itself in the language and script (wherever practicable) in which it appears there:
>
> Title and statement of responsibility
> Edition
> Publication, distribution, etc.
> Series
>
> Replace symbols or other matter that cannot be reproduced by the typographical facilities available with a cataloguer's description in square brackets. Make an explanatory note if necessary.

The main topic I want to examine is transcription in the new computing environment brought about by the Unicode Standard[2] and International Standard ISO/IEC 10646.[3] These publications cover not just the writing systems for all the major languages of the world, but collections of symbols and other elements of text, e.g., mathematical operators, Braille, punctuation, "dingbats", etc. Great care is taken to keep their character repertoires synchronized.

I also want to examine the issue of faithful transcription, what I call "exactitude" of cataloguing. Throughout, I will mention effects on retrieval, especially intersystem searching, which we must bear in mind as we make cataloguing decisions.

Now it was possible to have automated support for multiple scripts before the Unicode Standard and ISO/IEC 10646 - RLIN's scripts started with CJK in 1983,[4] and East Asian standards have always included several scripts - but with the availability of Unicode-based products, multiscript implementation is easier.

The Unicode Standard and ISO/IEC 10646 provide a much larger repertoire of scripts and characters than are currently authorized for any library application, including USMARC[5] and UNIMARC.[6] The expansion of the script repertoire means not only access to scripts that you never had before, but more characters for existing scripts. Here is a comparison for characters in several scripts.

Script	Character Category	USMARC/ UNIMARC	JIS X 0208 [7]	Unicode Standard Version 3.0
Cyrillic	Letters	102	66	237
Latin	Additional unaccented letters	21	0	163
Arabic	Letters	124	none	141
East Asian ideographs	Ideographs	13,469 (86% of EACC [8])	6,353	27,484

But please don't assume that the Unicode Standard and ISO/IEC 10646 will do everything for transcription:

1. Not everything that you may see on a source of information is **in** their repertoires.
2. Not everything you think you need for transcription **can be** in their repertoires.
3. Certain scripts require additional implementation support and extended fonts for correct presentation.

Which is not to say that you should reject these standards - I just want you to understand reality.

What's not there

The good news is that, with the addition of Sinhala, Ethiopic and Mongolian, all the major scripts of the world are now encoded. Version 3.0 of the Unicode Standard is to be published later this year, and the second edition of ISO/IEC 10646 is scheduled for next year.

Growth of the repertoire has not ended: various scripts for minority languages are still outstanding, more symbols could be added, and significant extinct scripts such as hieroglyphics and cuneiform are pending. (There may not be many libraries which collect and catalogue papyri and clay tablets, but the extinct scripts are significant for scholarship in general and certain museums in particular.

A single font for even the current Unicode character repertoire would be very large, and it's more practical to have fonts only for the scripts your library has in its collections. What is more likely to occur as you catalogue is not lack of a script, but lack of a particular character, e.g., if the title of a work on mathematics includes a symbol that isn't in the Mathematical Operators block. So occasionally you can't transcribe 100% of what is on the source of information.

But, you protest, I thought the Universal Character Set would have everything that I could possibly need! The response is no, for various reasons.

- The thing that you see on the source of information is an extremely rare character, so was simply missed;
- The thing that you see is known, and is being studied for possible addition;
- The thing that you see is known, but is not regarded as a character according to the Unicode design principles.

Two Unicode design principles are particularly significant in determining what should be encoded as a character: *Characters, not glyphs* and *Unification across languages*. In addition, the Unified Repertoire and Ordering of Han ideographs ("Unified Han"), developed by the Ideographic Rapporteur Group, has rules which determine uniqueness for ideographs.

Characters, not glyphs means that some high-level typographical aspects are not significant

when it comes to determining the character repertoire. Examples of typographical aspects are:

- The *nashki* style of Arabic writing versus the *nastaliq* style;
- Different ways of writing an East Asian ideograph;
- Different ways of writing a Cyrillic letter in particular languages;
- Contractions, typographical digraphs, etc.

Unification across languages means that:

- The graphemes used to write a particular language (e.g., an alphabet) are not separately encoded;
- Different language-based ways of writing a letter or ideograph are not encoded as separate characters.

These design principles and rules determine what is to be uniquely encoded. And as a result, not everything that appears on a source of information is eligible to be directly encoded as a defined character. This limitation on what can be encoded directly as defined characters is not a failure of the Unicode Standard. It comes about because of a different and more sophisticated vision of what should be encoded in a character set.

The original approach to the representation of text in machine-readable form was to give a unique code to each discrete mark on paper, although there was unification for generally accepted cases (the lower case forms of Latin letters a and g, for example). Character sets for East Asian languages assigned individual codes to different ways of writing what is fundamentally the same ideograph. Library character sets generally exhibit this "encode what you see" approach too, except for the use of non-spacing marks to encode accented Latin letters, where a letter with a diacritical mark is encoded as two characters. (Critics would say the letter is "broken apart.")

The Unicode Standard introduced a layered approach to the representation of text. "The design for a character set encoding must provide precisely the set of code elements that allows programmers to design applications capable of implementing a variety of text processes in the desired languages."[2] One result is that the characters in encoded text do not necessarily correspond 1:1 with the elements of that text in eye-readable form.

The simplest type of text representation is *plain text* a pure sequence of character codes. Unicode data is plain text. But to render what is wanted exactly, it may be necessary to use higher level protocols, such as language identification or layout instructions, to produce *fancy text* or *rich text* . USMARC and UNIMARC also use only plain text, but their character sets may provide separate encodings for things that are unified in Unicode/ISO 10646.

So we need to consider these issues:

- How exact must we be in transcription?
- If we have to be ultra-exact, how can we achieve this when we use Unicode/10646?

Evaluation of exactitude of transcription

So this brings us to consider the issue of exactitude of transcription. How exact does transcription have to be? Why? What exceptions do we make (perhaps without conscious decision-making)? What "work-arounds" do we use when we don't have the necessary typographical facilities?

We need exactitude in transcription in order to represent the item being identified uniquely and so make it accessible. Notice, however, that we don't always transcribe the information from the item with 100% fidelity.

One reason for the lack of fidelity is that cataloguing rules or the interpretation of them by a cataloguing agency do not always require, and sometimes do not allow, specific data to be transcribed. Here's an example. The Hebrew language is normally written unvocalized, that is, without vowel points and other marks of pronunciation. But sometimes these pronunciation guides are printed on the source of information; for example, when the author or publisher wants a word to be pronounced in an uncommon way. The Library of Congress, in its guidelines for cataloguing Hebrew,[10] builds on Rule 1.0G, *Accents and other diacritical marks*, and interprets it (incorrectly, in my opinion) as forbidding the transcription of vocalization marks that appear on the source of information.

One exception to exactitude is necessitated by lack of typographical facilities, a problem recognized in Rule 1.0E. The solution that this rule allows is the description of an unavailable textual element. This introduces an issue for intersystem searching - should the interpolation be ignored in searching, or treated as a "wild card" that matches anything, or...? The user cannot be expected to know the exact description written by the cataloguer.

There are also unwritten rules for exceptions to exactitude. Except for antiquarian and other precious books, we routinely ignore font features, calligraphy, etc. when transcribing, without any attempt to note such features. This is based on practicality, since for most modern works, distinctions at a very detailed level aren't needed.

When typographical facilities for a whole script are lacking, there are various options. When the language of cataloguing uses Latin script, the chosen solution is often romanization: transliteration or transcription into a Latin script form of the original text. Wellisch[11] reported in 1976 that the LC romanization tables (now ALA/LC) were most widely used, followed by those of ISO. When the language of cataloguing is Russian or another language written in Cyrillic script, cyrillicization is sometimes done. But not all languages use an alphabet or syllabary, and other solutions are to translate the information into the local language, or maintain card catalogues by script.

Access is impeded by all these alternatives. Where a library uses romanization or cyrillicization, the searcher must know that fact, know which conversion scheme is used for a particular language, and be able to apply that scheme correctly to create a search argument. A searcher may not know about the library's practice and use a completely different scheme. For translations, the searcher's translation may not match that of the cataloger. Card catalogs, unless they have been published in book form, cannot be searched remotely.

Lack of coded characters?

These problems will be alleviated considerably through the introduction of Unicode/ISO 10646 into USMARC and UNIMARC. But the use of a greatly expanded script repertoire does not mean that everything may be transcribed exactly. I now want to look at situations where even Unicode/ISO 10646 won't bring about 100% fidelity.

Historically, a primary reason for exactitude in transcription was to provide a surrogate of the bibliographic entity with as much detail as possible. The detail was needed because we had no other way to present the item in a card or book catalogue.

Problems of exact transcription are usually pointed out for ideographs, but this is not exclusively the case. If you're cataloguing a sound recording, what do you do about the name symbol used by "the artist formerly known as Prince"?

One source of difficulty is mathematics, where 2-dimensional formulas must be forced into a 1-dimensional field. Sargent has described how to represent mathematical formulas using Unicode.

Problems with ideographs arise because either the ideograph is not yet encoded, or when variant forms of an ideograph are represented by a single coded value (as noted by Zhang & Zhen).[12] Unavailable ideographs include both truly unique ideographs (used for personal names) and those in common use in a particular environment but is not yet in Unified Han (e.g., some of the government-sanctioned ideographs used in Hong Kong, or ideographs occurring in geographic names). In this situation:

- The geta symbol can be substituted for the unavailable ideograph. The geta comes from Japanese typography and is a placeholder for an ideograph to be supplied later. The technique is used in USMARC records.
- Ideographic description characters are intended to help the user visualize the unavailable character. Version 3.0 of The Unicode Standard and the second edition of ISO/IEC 10646 include these characters.

When a particular typographic form has been unified with others, yet the cataloger wants to use only that particular form, these are possible solutions.

- Use a higher level protocol, e.g., SGML[13] mark-up, to insist that this character be presented in a particular writing style. (Since both USMARC and UNIMARC use plain text, this option is outside their current scope.)
- Present the ideographic data in the record using a font determined by the language and country codes in the record. For example, if the language code was chi and the code for the country of publication was cc, the font would be a simplified Chinese style. If the language code was jpn, the font should be one with typical kanji. (This option will only work where the coded information is unequivocal, and when the ideographs appearing on the work are consistent with the preferred form for the language of the work and place of publication.)
- The Unicode Technical Committee has been considering a proposal by which ideographic variants can be indicated in plain text. Perhaps this will provide a solution.

Preferred regional or language forms are not exclusive to ideographs. When the Urdu language is written in Arabic script, it is conventionally printed in the nastaliq style. The Arabic language is usually printed in the *nashki* style. (*Nashki* is the style of the font used in RLIN's implementation of Arabic script.) Since all of the information on the work will normally be in the same typographic style, this can be highlighted through a note, when the typographic style on the item is not the same as that of the system. This is a situation similar to the black letter and Fraktur typographic styles of European printing.

A general solution to the problem of inexact transcription in bibliographic records is to use hyperlinking. In a Web-based catalog, we can have a link to a picture (scanned image) of the actual source of information. The disadvantage of a scanned image is that it cannot be searched for a specific occurrence of a particular glyphic form, but this is an operation that is more likely to be applied to full text than to cataloging.

Conclusion

The editors of cataloguing rules should review the rules on transcription to determine whether changes are needed due to the new technical environment. The new technical environment includes not only use of Unicode/ISO 10646 but also the ability to search remote catalogues via Z39.50.

Those in charge of the various MARC formats have to work with cataloguers to determine whether it is necessary to re-evaluate the "plain text" of the current formats. It isn't just a case of declaring Unicode/ISO 10646 as an approved character set (as has been done for UNIMARC[14]) or specifying the necessary changes in detail (as is underway for both

USMARC[15] and UNIMARC). That is the first and essential step, but cataloguing requirements may call for something beyond the "plain text" of the Unicode Standard and ISO/IEC 10646. If this is a requirement, then the various MARC formats will need to specify a methodology to provide this.

The question that has to be answered is: Is cataloguing data "plain text" or does it need to be a little fancier?

References

1 Anglo-American Cataloguing Rules, prepared under the direction of the Joint Steering Committee for Revision of AACR2; edited by Michael Gorman and Paul W. Winkler. 2nd ed., 1988 revision. (Chicago: American Library Association, 1988).

2 The Unicode Standard, Version 2.1 consists of:

- The Unicode Consortium, The Unicode Standard, Version 2.0, Addison-Wesley, Reading, MA, 1996. (ISBN 0-201-48345-9)
- The Unicode Standard, Version 2.1. (Unicode Technical Report # 8) Published on the Web at http://www.unicode.org/unicode/reports/tr8.html
  For a paper copy of Version 2.1, contact the Unicode Consortium.
  Version 2.0 can be ordered online from the Unicode Consortium (specify delivery method) at http://www.unicode.org/unicode/uni2book/u2ord.html or through the book trade.

Unicode is a trademark of Unicode, Inc. and may be registered in some jurisdictions.

3 International Organization for Standardization. Information Technology -- Universal Multiple-Octet Coded Character Set (UCS), Part 1: Architecture and Basic Multilingual Plane, Geneva, 1993. (ISO/IEC 10646-1:1993).

This International Standard is augmented by Technical Corrigendum 1:1996, Technical Corrigendum 2:1998, and nineteen Amendments (published between 1996 and 1999).

4 RLG East Asian Studies Community. http://www.rlg.org/eas/index.html

5 USMARC Specifications for Record Structure, Character Sets, and Exchange Media, prepared by Network Development and MARC Standards Office, 1994 ed., Cataloging Distribution Service, Library of Congress, Washington, D.C, 1994.

USMARC Format for Bibliographic Data, including Guidelines for Content Designation, prepared by Network Development and MARC Standards Office, 1994 ed., Cataloging Distribution Service, Library of Congress, Washington, D.C, 1994 -

USMARC Format for Authority Data, including Guidelines for Content Designation, prepared by Network Development and MARC Standards Office, 1993 ed., Cataloging Distribution Service, Library of Congress, Washington, D.C, 1993 -

For additional USMARC documentation see the Library of Congress' Web site.

6 UNIMARC Manual: Bibliographic Format, B. P. Holt and S. H. McCallum, eds., 2d ed., Saur, Munich, 1994.

UNIMARC/Authorities: Universal Format for Authorities, Saur, Munchen, 1991. (ISBN 3-598-10986-5)

7 Japanese Standards Association. Code of the Japanese Graphic Character Set for Information

Interchange. [English translation of JIS X 0208-1983] Tokyo, 1987. (JIS X 0208-1983)

8 American National Standards Institute, East Asian Character Code for Bibliographic Use, Transaction, New Brunswick, NJ, 1990. (ANSI Z39.64-1989).

9 The Unicode Standard, Version 2.0, p. 2-2.

10 Library of Congress. Descriptive Cataloging Division. Hebraica Cataloging: a guide to ALA/LC Romanization and Descriptive Cataloging, prepared by Paul Maher (Descriptive Cataloging Division). Cataloging Distribution Service, Library of Congress, Washington, D.C, 1987.

11 Wellisch, Hans H., "Script Conversion Practices in the World's Libraries," International Library Review 8:55-84 (1976).

12 Zhang, Foster J. and Zeng, Marcia Lei , Multiscript information processing on crossroads: demands for shifting from diverse character code sets to the Unicode Standard in library applications (Paper at 64th IFLA General Conference, 1998)
http://www.ifla.org/IV/ifla64/058-86e.htm

13 International Organization for Standardization. Information processing -- Text and office systems -- Standard Generalized Markup Language (SGML), Geneva, 1986. (ISO 8879:1986)

This International Standard is augmented by Technical Corrigendum 1:1996 and Amendment 1:1988.

14 UNIMARC Manual: Bibliographic Format, 2d. ed., Update 2 (1998).

15 Unicode Identification and Encoding in USMARC Records, submitted by MARBI Unicode Encoding and Recognition Technical Issues Task Force, 1998. (MARBI Proposal No: 98-18)
http://lcweb.loc.gov/marc/marbi/1998/98-18.html

Latest Revision: *June 9, 1999*

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").